The Geometry of Dynamic Scenes - On Coplanar and Convergent Linear Motions Embedded in 3D Static Scenes

Adrien Bartoli

The Geometry of Dynamic Scenes — On Coplanar and Convergent Linear Motions Embedded in 3D Static Scenes

Adrien Bartoli

University of Oxford
Department of Engineering Science
Ewert House, Ewert Place
Summertown
Oxford OX2 7BZ
United Kingdom

e-mail: *Bartoli@robots.ox.ac.uk*

Running head: Coplanar and Convergent Linear Motions

Revised Version, 17 dec. 2003

Corresponding author: *Adrien Bartoli*
Tel/fax: +44 1865 280 947 / 922

Keywords: Dynamic Scene, Structure From Motion, Matching Tensors.

Paper accepted in *Computer Vision and Image Understanding*

Abstract

In this paper, we consider structure and motion recovery for scenes consisting of static and dynamic features. More particularly, we consider a single moving uncalibrated camera observing a scene consisting of points moving along straight lines converging to a unique point and lying on a motion plane. This scenario may describe a roadway observed by a moving camera whose motion is unknown.

We show that there exist matching tensors similar to fundamental matrices. We derive the link between dynamic and static structure and motion and show how the equation of the motion plane (or equivalently the plane homographies it induces between images) may be recovered from dynamic features only.

Experimental results on real images are provided, in particular on a 60-frames video sequence.

1 Introduction

Most existing works on the geometry of multiple images rely on the assumption that the observed scene is rigid. The rigidity constraint allows to derive matching relations among two or more images, represented by e.g. the fundamental matrix or trifocal tensors. These matching tensors encapsulate the motion and the intrinsic parameters of the cameras which took the underlying images, and thus all the geometric information needed to perform 3D reconstruction. Matching tensors for rigid scenes can also be employed for scenes composed of multiple, independently moving objects [3, 5, 16], which requires however that enough features be extracted for each object, making segmentation, at least implicitly, possible.

On the other hand, there is a growing body of literature [1, 6, 7, 10, 11, 15, 17] dealing with the case of independently moving features, often termed as dynamic features. The goal of these works is to provide algorithms for dynamic structure and motion recovery as well as matching tensors for images of dynamic features. General, as well as highly constrained, dynamic scenarios, involving monocular or stereo views, have been investigated.

In this paper, we consider that the observed scene has both a static and a dynamic part. The static part is unconstrained (but has to be 3D) whereas on the other hand, as in [1, 6, 7, 11, 15, 17], we consider that dynamic features move along straight lines, termed *motion lines*. To further constrain the scenario, we consider that all motion lines lie on a motion plane and converge to an incidence point. Figure 2 illustrates this setting. Note that no assumption is made about the camera motion, which rules out background subtraction techniques, and that the camera is not assumed to be calibrated. A real-world instance of this scenario may be the motion of points arising from roadways seen from above, as for instance, by a moving surveillance video camera, see figure 1.

This scenario fits into less constrained cases previously examined [1, 6, 7, 11, 15, 17]. The corresponding dynamic structure-and-motion algorithms and matching tensors may therefore be used. The main drawback is that they require, in general, a number of point correspondences that may not be well-adapted for e.g. robust estima-

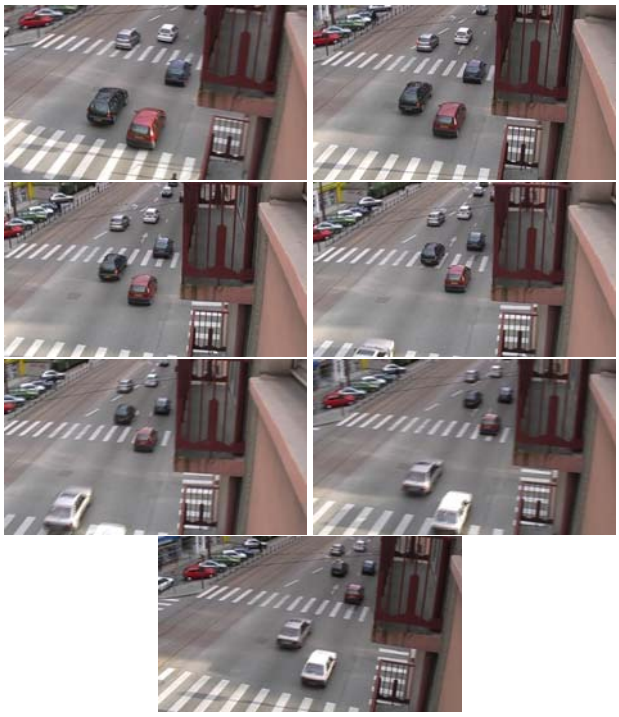

Figure 1: Selected frames of a video sequence consisting of 60 frames. Modeling the geometry of images provided by a moving camera observing a roadway is one application of our scenario. The difficulties of handling such sequences come from the fact that the scene consists of both static and dynamic features and that there is no constraint on the unknown camera motion.

tion based on random sampling techniques, that is most of the time required to devise practical systems.

Moreover, the following drawbacks arise. The method proposed in [1] requires that the camera motion is known and that point correspondences over 5 images are provided. The solutions proposed in [6, 7] rely on the fact that observed features have constant velocity, as well as most applications, apart from segmentation tensors, provided in [15]. In [11, 17], 3D views of the scene are required, which implies the use of two or more synchronized cameras. The H-tensor of [10] necessitates at least 3 images of static or dynamic points to be computed.

We show that much simpler matching tensors and dynamic structure and motion algorithms may be derived for the case studied in this paper.

Firstly, in §3, we examine the purely dynamic case, i.e. when only dynamic points are observed and when camera motion is unknown. We show that in the two-view case, dynamic structure and motion may be described by a fundamental matrix-like tensor that we call the C-tensor. Standard techniques, such as robust estimation [13] and maximum likelihood estimation through bundle adjustment [14] can then be applied in a straightforward manner to recover this tensor. We then show how dynamic motion in the multiple-view case can be modeled using a network of constrained C-tensors. A means to consistently estimate this geometry is provided. Dynamic structure in this case is also examined.

Secondly, in §4, we investigate the links between the previously-derived dynamic structure and motion and the projective registration (i.e. static motion) of the images. We give means for constrained estimation of camera motion.

Thirdly, in §5, we derive a matching tensor that is valid for both static and dynamic points. Prior segmentation of points into static/dynamic is therefore not necessary. This tensor is based on [16] and is modeled by a (6×6) matrix.

Experimental results on real images may be seen throughout the paper and in §6.

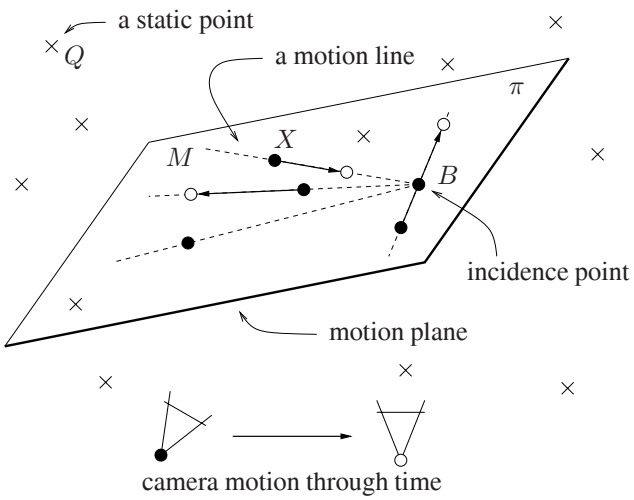

Figure 2: The dynamic scenario considered in this paper — unconstrained static points and points moving independently in a pencil of lines converging to the incidence point and lying on the motion plane. The motion plane and the incidence point are assumed to be static. However, most results derived in this paper are still valid without this assumption. In most practical cases, the incidence point lies at infinity since motion lines are parallel.

2 Background and Notation

We consider sets of 3D points, each of them denoted as U, that may be split into dynamic and static points respectively denoted as X and Q. Corresponding time-varying 3D coordinates are respectively denoted by $\mathbf{U}, \mathbf{U}', \mathbf{U}'', \ldots$, $\mathbf{X}, \mathbf{X}', \mathbf{X}'', \ldots$ and $\mathbf{Q}$. Images of these points are respectively denoted by $\mathbf{u}, \mathbf{u}', \mathbf{u}'', \ldots$, $\mathbf{x}, \mathbf{x}', \mathbf{x}'', \ldots$ and $\mathbf{q}, \mathbf{q}', \mathbf{q}'', \ldots$. Figure 2 illustrates some of these notations. The incidence point is denoted by B. It has coordinates $\mathbf{B}$ and projects to $\mathbf{b}, \mathbf{b}', \mathbf{b}'', \ldots$. It lies on the motion plane π that has coordinates $\boldsymbol{\pi}^\mathsf{T} \sim (\bar{\boldsymbol{\pi}}\ 1)$. The projective space of dimension d is denoted by $\mathbb{P}^d$. Everything is homogeneous (i.e. defined up to scale).

3 Purely Dynamic Views

Here we restrict to the case where only dynamic points can be observed from the scene. We assume that the different views are not registered, i.e. projection matrices are not available. We derive dynamic matching tensors for the two- then the multiple-view case. Figure 3 shows a toy example overlaid with dynamic features.

3.1 Two Views: The 7-dof C-Tensor

Figure 3: Dynamic points used for the experiments on the toy images. Note that four points are lying on a car which overtakes another one in the second image, and therefore do not fulfill the dynamic motion associated to the other points.

Derivation. We propose a way to derive the C-tensor, encapsulating the dynamic two-view motion. Alternatively, other means could be used, such as $\mathbb{P}^3 \rightarrow \mathbb{P}^2$ projec-

tion matrices within the framework of [15] or similarly to the join tensors of [17].

Let H be the unknown homography induced by the motion plane between the two views considered. Using H, we may predict the projection $\widetilde{\mathbf{x}}'$ of X in the second view if X was static as $\widetilde{\mathbf{x}}' \sim \mathsf{H}\mathbf{x}$. The image line $\mathbf{m}'$ of the motion line associated to X can then be obtained in the second view as the line joining $\mathbf{b}'$ and $\widetilde{\mathbf{x}}'$:

$$\mathbf{m}' \sim [\mathbf{b}']_\times \mathsf{H}\mathbf{x},$$

where $[\mathbf{b}']_\times$ is the (3×3) skew-symmetric cross-product matrix. Obviously, a necessary condition for $\mathbf{X}$ and $\mathbf{X}'$ to be instances of the same dynamic point X is that $\mathbf{x}'$ lies on $\mathbf{m}'$, which yields:

$$\mathbf{x}'^{\mathsf{T}} \mathcal{C} \mathbf{x} \text{ with } \mathcal{C} \sim [\mathbf{b}']_\times \mathsf{H}, \tag{1}$$

where we call $\mathcal{C}$ the 7-dof C-tensor (see below). It encapsulates the image signature of the dynamic two-view motion for the scenario previously described.

It is straightforward to see that the C-tensor has the same algebraic structure as the fundamental matrix. More precisely, the following analogy may be established. The projections of the incidence point play the roles of the epipoles while the 1D homography between the two motion line pencils corresponds to the epipolar transformation.

Properties. From the above-proposed analogy, several properties of the C-tensor may be easily derived. The C-tensor is rank-2 and has 7 dof. The projection of the indidence point in the first image, respectively in the second image, is the right null-space, respectively the left null-space, of the C-tensor: $\mathcal{C}\mathbf{b} = \mathcal{C}^{\mathsf{T}}\mathbf{b}' = \mathbf{0}_{(3\times1)}$. The extended motion line transformation G (a 5-dof 2D line-to-line homography relating the motion line pencils) is linked to the C-tensor as:

$$\mathsf{G} \sim \mathcal{C}\,[\mathbf{b}]_\times \text{ and } \mathcal{C} \sim \mathsf{G}\,[\mathbf{b}]_\times\,. \tag{2}$$

To understand the above expressions, consider a motion line $\mathbf{m}$ in the first image: $[\mathbf{b}]_\times\,\mathbf{m}$ is a point on this line ($\mathbf{b}$ is interpreted as a line that does not contain the point $\mathbf{b}$) and $\mathcal{C}\,[\mathbf{b}]_\times$ is the corresponding motion line in the other image. A similar reasoning may be done to understand the expression of $\mathcal{C}$ from G.

Similarly, a canonic plane homography, denoted as $\mathsf{H}^\star$ can be recovered as well as a 3-dimensional set of 2D homographies $\mathsf{H}_\mathbf{a}$ consistent with $\mathcal{C}$:

$$\mathsf{H}^\star \sim [\mathbf{b}']_\times \mathcal{C} \text{ and } \mathsf{H}_\mathbf{a} \sim \mathsf{H}^\star + \mathbf{b}'\mathbf{a}^\mathsf{T}. \tag{3}$$

Note that $\mathsf{H}_\mathbf{a}$ is a set of 2D homographies containing the plane homography H induced by the motion plane. We will see in §4.1 that when the fundamental matrix F (weak calibration of the cameras) is available, it is possible to recover the unknown H by computing the intersection of the family $\mathsf{H}_\mathbf{a}$ and the 3-dimensional family of plane homographies defined by F.

Estimation. Another consequence of the analogy between the C-tensor and the fundamental matrix is that one can apply any two-view projective structure and motion algorithm to estimate $\mathcal{C}$. For instance, we use the 8 point algorithm [9] embedded in a RANSAC-based robust estimation scheme [4] to compute an initial guess of $\mathcal{C}$, that we further refine using uncalibrated two-view bundle adjustment. In this case, the projective depths of points represent in fact their displacement along the motion lines. Figure 4 shows the result of computing pair-wise C-tensors.

Figure 4: Motion line pencils estimated using two-view projective bundle adjustment. For the middle image, we compute the motion line pencils with respect to two C-tensors (with the first and with the third images). Note the significant discrepancy between them. This discrepancy will be eliminated by a consistent parameterization of the multiple-view case, see §3.2. Note also that points lying on the overtaking car have been discarded as outiers.

3.2 Multiple Views: The 5-dof C-Tensor

We show that the relationships between n unregistered dynamic views are contained in $n-1$ C-tensors submitted to some additional consistency constraints, which express the fact that the incidence point is unique, regardless of the current time instant.

Degrees-of-freedom analysis and derivation. From the previous section, we already know that for $n = 2$ views, the dynamic geometry has 7 dof and is represented by a 7-dof C-tensor, say $\mathcal{C}$. Consider now a third view of the same scene which shares dynamic features with at least one of the two other views, say the second one. One can compute the C-tensor $\mathcal{C}'$ between the second and the third view. However, one has to remember that for a given time instant, the incidence point has a fixed position. Therefore, $\mathcal{C}$ and $\mathcal{C}'$ have to share the second image $\mathbf{b}'$ of the incidence point for being consistent, which provides 2 constraints and leaves $7 + 7 - 2 = 12$ dof for the dynamic geometry for $n = 3$. It is then straightforward to derive that the n-view case has $7 + 5(n-2) = 5n - 3$ dof.

The dynamic geometry of a set of n images can therefore be conveniently modeled by a 7-dof C-tensor between two reference views and a network of $n-2$ 5-dof C-tensors. A 5-dof C-tensor is a C-tensor with one constrained incidence point. A means to compute such a constrained C-tensor, given its right kernel, is provided in the next paragraph.

Concerning the dynamic structure, each point has $2 + (n-1) = n + 1$ dof corresponding to its position in the motion plane and $n-1$ motions along its motion line.

Threading C-tensors. Once a solution has been obtained for the 7-dof C-tensor modeling the dynamic geometry of two particular images, subsequent 5-dof C-tensors have to be computed given one projection of the incidence point, e.g. with their left or right kernel known. Enforcing these consistency constraints when threading C-tensors is important since only 5 point correspondences instead of 7 are necessary to solve for the constrained tensor, as shown below. Moreover, the solution obtained is consistent and may be refined directly without prior correction using non-linear methods to obtain, e.g. a maximum likelihood solution.

We propose a linear algorithm inspired by [8] for estimating 5-dof C-tensors using 5 or more point correspondences while enforcing the consistency constraints. Using equation (1) and the factorization (2) of the C-tensor, we may write:

$$\mathbf{x}'^{\mathsf{T}} \mathsf{G} \left[\mathbf{b}\right]_{\times} \mathbf{x} = 0, \tag{4}$$

where G contains the unknown motion line pencil transformation and $\mathbf{b}$ is the known projection of the incidence point in the first image. Let us see how to solve for G. Let $\left[\mathbf{b}\right]_{\times} \sim \mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}}$ be a singular value decomposition of $\left[\mathbf{b}\right]_{\times}$ where $\Sigma = \mathrm{diag}(1, \sigma, 0)$. An efficient solution to obtain this decomposition is given in [8]. By replacing into equation (4), we obtain:

$$\mathbf{x}'^{\mathsf{T}} \bar{\mathsf{G}} \bar{\mathbf{y}} \quad = \quad 0 \text{ with } \begin{pmatrix} & & 0 \\ & \bar{\mathsf{G}}_{(3\times 2)} & 0 \\ & & 0 \end{pmatrix} = \mathsf{G}\mathsf{U}\Sigma \text{ and } \begin{pmatrix} \bar{\mathbf{y}}_{(2\times 1)} \\ y \end{pmatrix} = \mathsf{V}^{\mathsf{T}}\mathbf{x}.$$

Note that $\bar{\mathsf{G}}$ is defined by 6 homogeneous parameters $\bar{\mathbf{g}} \in \mathbb{P}^5$, which is consistent with the fact that G has 5 dof[1]. Expanding this equation leads to the following homogeneous linear system for $\bar{\mathbf{g}}$:

$$\mathsf{A}_{(m\times 6)} \cdot \bar{\mathbf{g}}_{(6\times 1)} = \mathbf{0}_{(m\times 1)} \text{ with } \mathsf{A} = \begin{pmatrix} & & \cdots & & & \\ x'_1\bar{y_1} & x'_1\bar{y_2} & x'_2\bar{y_1} & x'_2\bar{y_2} & \bar{y_1} & \bar{y_2} \\ & & \cdots & & & \end{pmatrix},$$

where m is the number of point correpondences considered. Note that 5 equations are sufficient to solve for $\bar{\mathsf{G}}$ in the least squares sense. The singular vector associated with the smallest singular value of A, that may be obtained using singular value decomposition, provides the least squares solution for $\bar{\mathbf{g}}$, then $\bar{\mathsf{G}}$. From $\bar{\mathsf{G}}$, one can further obtain $\mathcal{C}$ as:

$$\mathcal{C} \sim \begin{pmatrix} & & 0 \\ & \bar{\mathsf{G}}_{(3\times 2)} & 0 \\ & & 0 \end{pmatrix} \mathsf{V}^{\mathsf{T}}.$$

[1] $\bar{\mathbf{g}}$ is the row-wise vectorization of $\bar{\mathsf{G}}$.

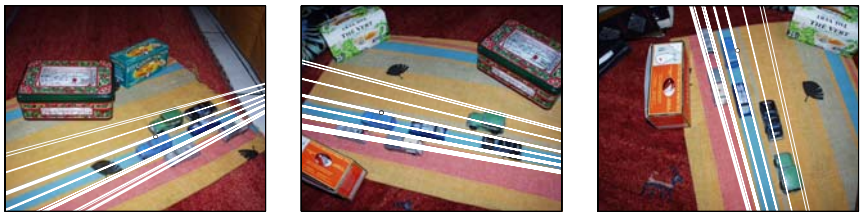

Figure 5: Motion line pencils estimated using two-view projective bundle adjustment between the first and the second views and constrained estimation of a 5-dof C-tensor between the second and third views. Note that the motion lines pencils are perfectly aligned in the second image.

Estimation. The 7-dof C-tensor between two particular views may be estimated using the 8 point algorithm. Other 5-dof C-tensors may be computed using the above-described constrained method, possibly embedded in a RANSAC-based robust estimation process [4]. Figure 5 shows the result of such a constrained computation. This provides an initial guess of the dynamic registration, which may be refined as follows.

As in the case of multiple-view bundle adjustment, minimizing the reprojection error, i.e. the discrepancy between measured and predicted features yields the maximum likelihood estimator. We employed such a means directly for the two-view case since there was a direct analogy. For the multiple-view case however, one can not directly use standard bundle adjustment techniques for the following reasons. Firstly, the projective multiple-view motion has $11n - 15$ dof whereas dynamic motion involves $5n - 3$ dof. Secondly, a reconstructed static point has 3 dof whereas a dynamic point has $n + 1$ dof (provided it is visible in every image considered). Therefore, the problem must be specifically parameterized.

A consistent parameterization of the problem is the following. Dynamic motion can be parameterized using n images of the incidence point and $n - 1$ 3-dof motion line pencil transformations, which yield the required $2n+3(n-1) = 5n-3$ dof. Minimally parameterizing these entities can be done using techniques inspired from standard non-linear estimation of the epipolar geometry, see e.g. [2, 18]. Dynamic structure can be parameterized as one unconstrained image point and $n - 1$ image points constrained

on known image lines, which yield the required $2 + n - 1 = n + 1$ dof. Representing an image point on a known image line can be done using 1 parameter.

4 Mixing Static and Dynamic Features

In this section, we consider that enough static points may be used to perform a weak calibration of the cameras, i.e. to projectively register them. A classical means for such a registration is to compute static structure and motion between two particular views and iteratively register the other views using 3D-to-2D point correspondences. We derive the links between the dynamic structure and motion given in §3 and classical static structure and motion and show how the above-mentioned registration algorithm can be constrained by known C-tensors.

4.1 Two Views

We represent the projective two-view motion by the fundamental matrix F. Firstly, we investigate the link between F and the C-tensor $\mathcal{C}$. Secondly, we show that if F and $\mathcal{C}$ are known, the plane homography H induced by the motion plane may be recovered and we give a closed-form solution in terms of F and $\mathcal{C}$ as well as a means to use standard homography estimation algorithms.

4.1.1 The Link Between $\mathcal{C}$ and F

To establish this link, we consider the plane homography H induced by the motion plane. This homography can be written in terms of the C-tensor as $\mathsf{H}_{\mathbf{a}}$, see equation (3), where the unknown 3-vector $\mathbf{a}$ is used to span the space of 2D homographies consistent with $\mathcal{C}$. The fundamental matrix can be formed from any plane homography as $\mathsf{F} \sim [\mathbf{e}']_\times \mathsf{H}$ and in particular $\mathsf{H}_{\mathbf{a}}$ which yields:

$$\mathsf{F} \sim [\mathbf{e}']_\times ([\mathbf{b}']_\times \mathcal{C} + \mathbf{b}'\mathbf{a}^\mathsf{T}). \tag{5}$$

This equation, that we call the F-$\mathcal{C}$*-consistency constraint*, shows that F has only 5 dof corresponding to the right epipole and the equation $\mathbf{a}$ of the motion plane. It is

equivalent to the fact that F and $\mathcal{C}$ share a 2D homography. Therefore, given $\mathcal{C}$, 5 point correspondences should be enough to estimate F. However, due to the non-linearity of equation (5) for the unknowns $\mathbf{e}'$ and $\mathbf{a}$, we can not estimate F linearly using 5 point correspondences. Another solution is to use the fact that the incidence point $\mathbf{B}$ is a static point and that therefore, $\mathbf{b}$ and $\mathbf{b}'$ give one constraint on F through the fundamental equation $\mathbf{b}'^{\mathsf{T}}\mathsf{F}\mathbf{b} = 0$. A minimum of 6 other static points are then sufficient to estimate F.

4.1.2 Retrieving H

Given the C-tensor and the epipolar geometry, it is possible to recover the plane homography induced by the motion plane π. The following three paragraphs give respectively an analysis of the generic degenerate cases, in which the motion plane can not be uniquely recovered, a closed-form solution in terms of the C-tensor and the fundamental matrix and a more physically meaningful solution taking feature positions into account.

Point prediction and degenerate cases. Let $\mathbf{x}, \widetilde{\mathbf{x}}'$ be the projections of a 3D point $\mathbf{X} \in \pi$ in two images. Point $\widetilde{\mathbf{x}}'$ can be predicted by intersecting the motion line and the epipolar line associated to $\mathbf{x}$ in the second image:

$$\widetilde{\mathbf{x}}' \sim (\mathcal{C}\mathbf{x}) \times (\mathsf{F}\mathbf{x}). \tag{6}$$

This *prediction equation* is used in the two methods proposed below for recovering the plane homography H. It is valid provided that $\mathbf{x}$ does not lie on the image line $\mathbf{b} \times \mathbf{e}$. Indeed, if $\mathbf{x} \in (\mathbf{b} \times \mathbf{e})$, then the motion line $\mathbf{x} \times \mathbf{b}$ is the same as the epipolar line $\mathbf{x} \times \mathbf{e}$, and the only constraint that one can infer on point $\widetilde{\mathbf{x}}'$ is that it lies on the line $\mathsf{F}\mathbf{x} \sim \mathcal{C}\mathbf{x}$, the epipolar / predicted motion line. Including this condition in the prediction equation (6) yields:

$$\forall \mathbf{x} \in \mathbb{P}^2,\ \left(\mathbf{x}^{\mathsf{T}}(\mathbf{b} \times \mathbf{e}) \neq 0\right) \Rightarrow \left(\widetilde{\mathbf{x}}' \sim (\mathcal{C}\mathbf{x}) \times (\mathsf{F}\mathbf{x})\right). \tag{7}$$

We now turn to studying the degenerate cases where point $\widetilde{\mathbf{x}}'$ can not be uniquely predicted whatever point $\mathbf{x}$ is, i.e. degenerate cases depending on the camera positions and the direction of motion. These degenerate cases are described by the fact

that the epipolar and the predicted motion line in the second image are identical, i.e. $\forall \mathbf{x} \in \mathbb{P}^2,\ \mathcal{C}\mathbf{x} \sim \mathsf{F}\mathbf{x}$. Obviously, this appends when $\mathcal{C} \sim \mathsf{F}$. We show below that this corresponds to the fact that the incidence point $\mathbf{B}$ lies on the baseline i.e. the line joining the two centers of projection $\mathbf{C}$ and $\mathbf{C}'$. We use the following equivalence: $(\mathbf{B} \in (\mathbf{C}\mathbf{C}')) \Leftrightarrow (\mathbf{e} \sim \mathbf{b}) \Leftrightarrow (\mathbf{e}' \sim \mathbf{b}')$. Replacing $\mathbf{e}'$ by $\mathbf{b}'$ in the F-$\mathcal{C}$-consistency constraint (5) yields:

$$\begin{aligned} \mathsf{F} &\sim [\mathbf{b}']_\times ([\mathbf{b}']_\times \mathcal{C} + \mathbf{b}'\mathbf{a}^\mathsf{T}) \\ &\sim [\mathbf{b}']_\times^2 \mathcal{C} \\ &\sim \mathcal{C}, \end{aligned}$$

which shows[2] $(\mathbf{B} \in (\mathbf{C}\mathbf{C}')) \Rightarrow (\mathsf{F} \sim \mathcal{C})$. Showing $(\mathsf{F} \sim \mathcal{C}) \Rightarrow (\mathbf{B} \in (\mathbf{C}\mathbf{C}'))$ is straightforward: $(\mathsf{F} \sim \mathcal{C}) \Rightarrow (\mathbf{e} \sim \mathbf{b}) \Rightarrow (\mathbf{B} \in (\mathbf{C}\mathbf{C}'))$. Hence, the prediction equation (6) degenerates if and only if the incidence point lies on the baseline. Note that the formulation (7) accounts for this case since $(\mathbf{B} \in (\mathbf{C}\mathbf{C}')) \Leftrightarrow (\mathbf{b} \sim \mathbf{e}) \Rightarrow (\forall \mathbf{x} \in \mathbb{P}^2,\ \mathbf{x}^\mathsf{T}(\mathbf{b} \times \mathbf{e}) = 0)$. This means that $\mathbf{x}^\mathsf{T}(\mathbf{b} \times \mathbf{e}) = 0$ describes all generic degenerate cases for point prediction or computation of the plane homography H.

In practice, when the incidence point is at infinity, a degeneracy occurs when the baseline is parallel to the motion lines. This configuration can easily be avoided, e.g. for road surveillance, if more that one camera are used, then they can be mounted on a bridge perpendicular to the road.

A closed-form solution. In general, the mapping represented by equation (7) is bilinear in $\mathbf{x}$, i.e. it does not correspond to an homography. We show that it reduces to a linear mapping, which is the plane homography H, provided that the F-$\mathcal{C}$ consistency constraint is satisfied, i.e. that they are compatible with a shared H. In this case, we may write $\mathcal{C} \sim \mathsf{H}^{-\mathsf{T}}[\mathbf{b}]_\times$ and $\mathsf{F} \sim \mathsf{H}^{-\mathsf{T}}[\mathbf{e}]_\times$, which yields $\widetilde{\mathbf{x}}' \sim \left[\mathsf{H}^{-\mathsf{T}}[\mathbf{b}]_\times \mathbf{x}\right]_\times \mathsf{H}^{-\mathsf{T}}[\mathbf{e}]_\times \mathbf{x}$. This equation reduces to $\widetilde{\mathbf{x}}' \sim \mathsf{H}\mathbf{x}$ after some algebraic manipulations and provided that $\mathbf{x}$ does not lie on the image line $\mathbf{b} \times \mathbf{e}$, which is a

[2] The last equality follows from the fact that $\mathcal{C}\mathbf{x}$ is the predicted motion line, $[\mathbf{b}']_\times \mathcal{C}\mathbf{x}$ is its intersection with the line of equation $\mathbf{b}'$ (which never contains the projection of the incidence point since $\mathbf{b}^\mathsf{T}\mathbf{b}' = \|\mathbf{b}'\|^2 \neq 0$), and finally, $[\mathbf{b}']_\times^2 \mathcal{C}\mathbf{x}$ is the predicted motion line itself, from which $[\mathbf{b}']_\times^2 \mathcal{C} \sim \mathcal{C}$ follows.

already required for the prediction equation to be valid.

We claim that H can be recovered as:

$$\mathsf{H} \sim (\mathcal{C} \times \mathsf{F}) \cdot \mathrm{diag}\left((\mathcal{C} \times \mathsf{F})^{-1} \left[\sum_i \mathbf{c}_i \right]_\times \left(\sum_i \mathbf{f}_i \right) \right), \tag{8}$$

where $\mathbf{c}_i$ and $\mathbf{f}_i$ designate the i-th column of $\mathcal{C}$ and F respectively, and $\mathcal{C} \times \mathsf{F}$ is the column-wise cross-product of $\mathcal{C}$ and F. The proof of this result is obtained as follows. We already know from equation (7) that $\forall \mathbf{x} \in \mathbb{P}^2$, $(\mathbf{x}^\mathsf{T}(\mathbf{b} \times \mathbf{e}) \neq 0) \Rightarrow (\mathsf{H}\mathbf{x} \sim (\mathcal{C}\mathbf{x}) \times (\mathsf{F}\mathbf{x}))$. Applying this equation to each of the 4 vectors forming the canonic basis of $\mathbb{P}^2$ yields the above-mentioned result. Figure 6 shows the result of estimating pair-wise homographies and predicting the position of the point features from one view to the others.

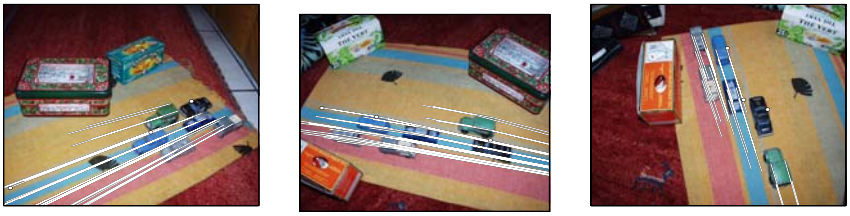

Figure 6: Dynamic points transferred between images using plane homographies associated with the motion plane, i.e. as if they were static. These plane homographies have been estimated by hallucinating point correspondences (see text) and minimizing the symmetric squared distance between measured and transferred features. The final error is 0.35 pixels while the initial guess provided by the closed-form solution (not shown here) has an error of 2.77 pixels.

Care must be taken about the use of equation (8). Indeed, when $\mathcal{C}$ and F are perfectly consistent, the resulting H is unique. On the other hand, when $\mathcal{C}$ and F have been independently estimated, equation (8) gives an approximated solution based on the transfer of the 4 vectors of the canonic basis of $\mathbb{P}^2$. Since two of these vectors represent points lying at infinity, the resulting homography may not be well-suited for points considered, usually lying in the images. Moreover, equation (8) is valid provided that the points represented by the vectors of the canonic basis of $\mathbb{P}^2$ do not lie

on the line $\mathbf{s} = (s_1\ s_2\ s_3)^{\mathsf{T}} \sim \mathbf{b} \times \mathbf{e}$ and hence, the following conditions apply: $s_1 \neq 0,\ s_2 \neq 0,\ s_3 \neq 0$ and $s_1 + s_2 + s_3 \neq 0$. These conditions mean in particular that the line $\mathbf{s}$ must not be parallel to the x-axis or to the y-axis and must not contain the origin of the image.

For these reasons, we do not recommend the above-described method for the estimation of the plane homography H from $\mathcal{C}$ and F, but rather the method described in the following paragraph.

A physically meaningful solution. More physically meaningful means to estimate H, but more computationally expensive, can be obtained by hallucinating static point correspondences that lie on the plane using equation (6). Any standard method can then be used to solve for H by minimizing a given criterion, see e.g. [9]:

1. hallucinate point correspondences as:

$$\left\{\cdots, (\mathbf{x}, (\mathcal{C}\mathbf{x}) \times (\mathsf{F}\mathbf{x})), \left(\mathbf{x}', (\mathcal{C}^{\mathsf{T}}\mathbf{x}') \times (\mathsf{F}^{\mathsf{T}}\mathbf{x}')\right), \cdots\right\};$$

2. use any standard method to estimate H.

For example, we have chosen to non-linearly optimize the following cost function using the Levenberg-Marquardt algorithm initialized by the previously-given closed-form solution:

$$\mathsf{H} \sim \arg\min_{\mathsf{H}} \sum_{\mathbf{x}\leftrightarrow\mathbf{x}'} \left(d^2\left(\mathsf{H}\mathbf{x}, (\mathcal{C}\mathbf{x}) \times (\mathsf{F}\mathbf{x})\right) + d^2\left(\mathsf{H}^{-1}\mathbf{x}', (\mathcal{C}^{\mathsf{T}}\mathbf{x}') \times (\mathsf{F}^{\mathsf{T}}\mathbf{x}')\right)\right).$$

Experimental results can be seen on figure 6. Note that hallucinated points are also used in [12] for the estimation of general structure and motion, given sets of coplanar points. This method is not subject to any degeneracy other than the generic degeneracies preventing the use of the prediction equation (7).

4.2 Multiple Views

In the registered case, 2 views entirely fix the dynamic motion, since the incidence point and the motion plane may be determined uniquely. On the other hand, each additional view of a dynamic point adds 1 dof, corresponding to its position on its motion line at the time instant the picture was taken, as in the purely dynamic case.

4.2.1 The Link Between $\mathcal{C}$ and P

Consider a set of images statically and dynamically registered. Without loss of generality, assume that two cameras are $\mathsf{P} \sim (\mathrm{I}_{(3\times3)}\ \mathbf{0}_{(3\times1)})$ and $\mathsf{P}' \sim (\mathsf{H}^{12}_{(3\times3)}\ \mathbf{e}'_{(3\times1)})$, where H^{ij} is the plane homography between view i and view j corresponding to the motion plane. We want to express the link between a third view with camera matrix P'' and a 5-dof C-tensor $\mathcal{C}$ describing the dynamic geometry between view 2 and view 3. With this choice for P and P', $\mathsf{P}'' \sim (\mathsf{H}^{13}_{(3\times3)}\ \mathbf{e}''_{(3\times1)})$, which may be written using equation (6) as:

$$\mathsf{P}'' \sim (\mathsf{H}^{12}([\mathbf{b}']_\times \mathcal{C} + \mathbf{b}'\mathbf{a}^\mathsf{T})\ \mathbf{e}''), \tag{9}$$

where $\mathbf{e}''$ is the epipole of view 1 into view 3. We call this equation the P-$\mathcal{C}$-*consistency constraint*.

4.2.2 Estimation.

Two cases may be considered. Estimation of the C-tensor given the projection matrix and conversely. We consider the latter case, altough there exist means for the former.

Using the P-$\mathcal{C}$-consistency constraint (9), one may write a linear system for the 6 unknowns $\mathbf{a}$ and $\mathbf{e}''$ where each 3D/2D static point correspondence $\mathbf{Q} \leftrightarrow \mathbf{q}$ gives 2 equations through $\mathbf{q} \sim \mathsf{P}\mathbf{Q}$. Therefore, 3 point correspondences are enough to solve for P'', instead of 6 in the unconstrained case.

It is possible to refine the obtained solution by non-linearly minimizing the reprojection error using techniques from bundle adjustment.

5 A Unified Tensor

The above-proposed methods share the same drawback. They require that static and dynamic points have been segmented. We derive a unified matching tensor for two views of both static and dynamic points, inspired from [16]. Consider a point correspondence $\mathbf{u}, \mathbf{u}'$ that may be either static or dynamic. If this point is static, then it must satisfy the fundamental equation $\mathbf{u}'^\mathsf{T}\mathsf{F}\mathbf{u} = 0$ and if it is dynamic, it must satisfy

equation (1). Therefore, the following constraint must hold:

$$\left(\mathbf{u'}^{\mathsf{T}}\mathsf{F}\mathbf{u}\right)\cdot\left(\mathbf{u'}^{\mathsf{T}}\mathcal{C}\mathbf{u}\right)=0.$$

It has been shown in [16] that expanding this equation, and after some algebraic manipulations, equation $\widehat{\mathbf{u}}'^{\mathsf{T}}\mathcal{S}\widehat{\mathbf{u}}=0$ can be obtained, where $\mathcal{S}$ is the (6×6) unified tensor (originally called "segmentation matrix") and $\widehat{\mathbf{u}}^{\mathsf{T}}\sim(u_1^2\ u_1u_2\ u_2^2\ u_1u_3\ u_2u_3\ u_3^2)$ are the 6 coordinates of $\mathbf{u}$ lifted onto $\mathbb{P}^5$. This tensor has nice properties to describe the geometry of the scenario considered. However, it is not well-suited for robust estimation since 35 point correspondences are required for its linear estimation. Moreover, a minimum of 8 static and dynamic point correspondences is required, which increases the number of iterations of e.g. a RANSAC procedure.

6 Experimental Results Using Real Images

We compute dynamic and static structure and motion on a 60-frame sequence from which sample images are shown on figure 1. We select dynamic and static features by hand on the first image and automatically track them through the sequence using a correlation-based technique. We then used key frames $0, 10, 20, \ldots$ shown on figure 1.

We first perform dynamic structure and motion by sequentially computing C-tensors as described in §3.2. We then perform constrained static structure and motion as described in §4. Lastly, we use these results to recover the plane homographies associated to the motion plane between key frames, as indicated in §4.1.2. Such homographies allow to transfer dynamic features and predict their position in another camera position and another time instant as if they were static. The result of such transfers is shown on figure 7.

Figure 7 also shows that the first homography (i.e. between frames 0 and 10) is relatively accurate since static point positions after transfer seem visually good.

The main problem that we encountered was the computation of the initial 7-dof C-tensor between the two first frames. Indeed, one may observe that all vehicles have roughly the same speed, which therefore induces a point-to-point homography between dynamic features of these two frames. There was therefore a 2-dof ambiguity on the

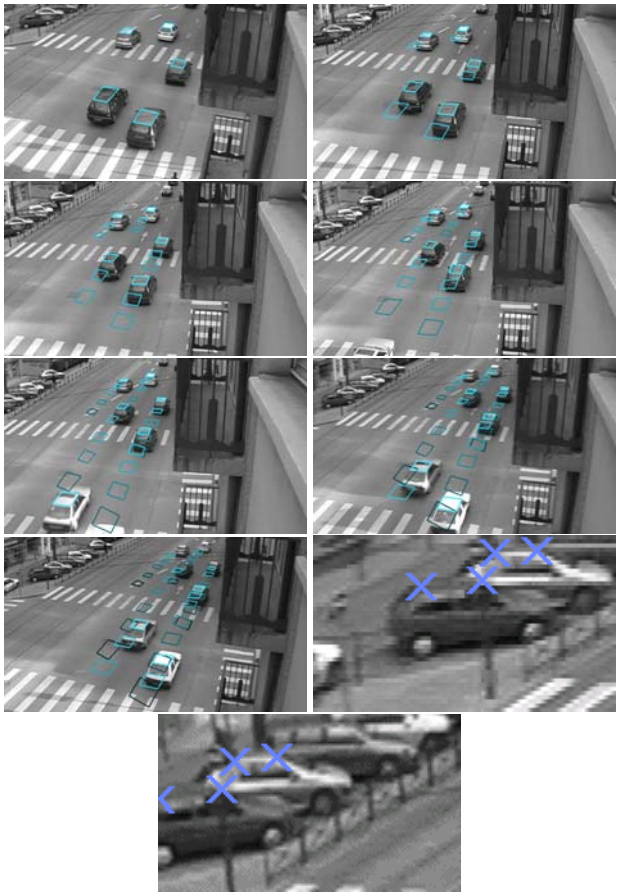

Figure 7: Plane homographies recovered using dynamic and static structure and motion allow to transfer vehicle positions from one key frame to the others as if they were static. The last two images are zooms on frames 0 and 10 respectively. The first one shows manually clicked static points (lying on the motion plane) while the second one shows the transfer of these points using the recovered plane homography (computed using dynamic points only).

computation of the 7-dof C-tensor. Instead, we computed a 5-dof C-tensor constrained by the projection of the incidence point in the first image. This projection was obtained by intersecting support lines of white bands on the ground.

7 Conclusion

We addressed the case of a specific dynamic scenario describing the motion of point features along lines converging to the same point and lying onto a motion plane. We show that very simple matching tensors that we call C-tensors, similar to fundamental matrices, exist. We show how to constrain static structure and motion by its dynamic counterpart. Plane homographies associated with the motion plane can then be recovered from dynamic features only. Experimental results show that this approach is feasible in practice and may be used to model e.g. surveillance video cameras observing roadways.

We believe that these geometrical features may be successfully used to devise completely automatic systems for vehicule tracking and camera motion estimation. Among issues for further work, self-calibration of the camera by considering that in practice the incidence lies most of the time at infinity, could be examined.

References

[1] S. Avidan and A. Shashua. Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 22(4):348–357, April 2000.

[2] A. Bartoli and P. Sturm. Three new algorithms for projective bundle adjustment with minimum parameters. Research Report 4236, INRIA, Grenoble, France, August 2001.

[3] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. In E. Grimson, editor, *Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA*, pages 1071–1076. IEEE, IEEE Computer Society Press, June 1995.

[4] M.A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis. Technical report, SRI International, Menlo Park, CA, 1980.

[5] A.W. Fitzgibbon and A. Zisserman. Multibody structure and motion: 3-D reconstruction of independently moving objects. In D. Vernon, editor, *Proceedings of the 6th European Conference on Computer Vision, Dublin, Ireland*, volume 1842 of *Lecture Notes in Computer Science*, pages 891–906. Springer-Verlag, June 2000.

[6] M. Han and T. Kanade. Reconstruction of a scene with multiple linearly moving objects. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, June 2000.

[7] M. Han and T. Kanade. Multiple motion scene reconstruction from uncalibrated views. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, July 2001.

[8] R. Hartley. Minimizing algebraic error. In *Proceedings of the 6th International Conference on Computer Vision, Bombay, India*, 1998.

[9] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, June 2000.

[10] A. Shashua and L. Wolf. Homography tensors: On algebraic entities that represent three views of static or moving planar points. In D. Vernon, editor, *Proceedings of the 6th European Conference on Computer Vision, Dublin, Ireland*, volume 1842 of *Lecture Notes in Computer Science*, pages 507–521. Springer-Verlag, June 2000.

[11] P. Sturm. Structure and motion for dynamic scenes – the case of points moving in planes. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, volume 2351 of *Lecture Notes in Computer Science*, pages 867–882, May 2002.

[12] R. Szeliski and P.H.S. Torr. Geometrically constrained structure from motion : Points on planes. In *3D Structure from Multiple Images of Large-scale Environments SMILE'98*. Springer Verlag, June 1998.

[13] P.H.S. Torr. *Motion Segmentation and Outlier Detection*. PhD thesis, University of Oxford, England, Department of Engineering Science, Parks Road, Oxford, 1995.

[14] B. Triggs, P.F. McLauchlan, R.I. Hartley, and A. Fitzgibbon. Bundle ajustment — a modern synthesis. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Proceedings of the Interna-*

tional Workshop on Vision Algorithms: Theory and Practice, Corfu, Greece, volume 1883 of *Lecture Notes in Computer Science*, pages 298–372. Springer-Verlag, 2000.

[15] L. Wolf and A. Shashua. On projection matrices $P^k \rightarrow P^2, k = 3, \cdots, 6$, and their applications in computer vision. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, volume 1, pages 412–419. IEEE Computer Society Press, July 2001.

[16] L. Wolf and A. Shashua. Two-body segmentation from two perspective views. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA*, December 2001.

[17] L. Wolf, A. Shashua, and Y. Wexler. Join tensors: On 3D-to-3D alignment of dynamic sets. In A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquézar, J.-O. Eklundh, and Y. Aloimonos, editors, *Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain*, volume 1, pages 388–391, September 2000.

[18] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–195, March 1998.